DOCUMENT RESUME

ED 115 671 TM 004 941

AUTHOR Donlon, Thomas F.
TITLE Estimating Rate-of-Work Parameters in the Assessment of Test Speededness.
PUB DATE [Apr 75]
NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Mathematical Models; *Statistical Analysis; *Timed Tests
IDENTIFIERS Power Tests; *Speededness (Tests); Variance (Statistical)

ABSTRACT

The use of indices of speededness in tests is reviewed and three possible indices which derive from a model which assumes that rate of work is normally distributed is proposed. Each of these indices is seen as limited by the failure to adequately consider the correlation between speed and power, but they have the advantage that they are derivable from a single administration of a test. The plausibility of an assumption of a normal distribution of work rates was tested on empirical data from seven tests, and in six cases the assumption was not unrealistic although the fit was only approximate. The analysis of tests for which the assumption is not completely tenable will often be instructive for those who construct instruments. Specific item material seems to govern departures from normality. The goal of a single, intuitively satisfying index of speededness in tests is important. Without some sort of metric or scale with which to assess degrees of speededness, the evaluation of tests in this important area remains inordinately subjective. (Author/BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115671

Estimating Rate-of-Work Parameters in the Assessment of Test Speededness

Thomas F. Donlon
Educational Testing Service
Princeton, New Jersey

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AERA - NCME
Washington, D. C.
1975

2

TM004 941

Estimating Rate-of-Work Parameters in the
Assessment of Test Speededness

The concept of test speededness is intuitively clear. If two individuals have about the same capacity for success, on the average, in dealing with the items in a test, but if one of them works much faster, then if the test is timed this faster individual will enjoy an advantage. A test is speeded, then to the extent that rate-of-work differences contribute to or explain the score variance. A speed test is one for which all score variance is thus explained, while a power test is one for which no score variance is thus explained.

Gulliksen (1950) has described a basic conception for a kind of scale for test speededness. His discussion considers a test in which all individuals have the same capacity for success, because they can all answer any question, but they have different work rates. Further, there is a stringent time limit. This situation is compared with a test on which the individuals differ markedly in their potential for correct answers, and in their rate of work, but there is no time limit. These models are referred to as a pure speed test and a pure power test.

The preponderance of tests in educational and psychological testing do not fit either of these models, but fall somewhere in between. Unfortunately, although the polar concepts of pure speed and pure power suggest a kind of scale by which to assess test speededness, no simple index has been developed for characterizing a test on this implicit continuum. This appears to result from the underlying complexity of the factors involved. Test speededness seems

logically related to a variety of such characteristics as the number of questions left unanswered, or answered wrongly, and a simple statistical formula has proved difficult to develop. Gulliksen suggests the examination of some variance ratios, considering the ratio of the variance of wrong answers to total score variance and the ratio of the variance of unattempted items to the total score variance. When the ratio of the variance of the wrong answers is small, the test may be considered primarily speeded. When the ratio of the variance of the unattempted items is small, the test may be considered predominantly a power test.

These ratios are consistent with the definitions, and they are calculable from a single test administration, but they are somewhat dissatisfying in that they only give strong results in extreme cases, when the observed values are very small. In many tests it is possible for both of the ratios to take quite sizable values, but there is no adequate technique for combining them into a single, logically consistent index, broadly applicable to the evaluation of tests. Further, as these variances of wrong answers and of unattempted items increase in magnitude the question of the correlation between wrong answers and unattempted items becomes a critical factor.

The salient role of this correlation is clear from a consideration of its impact under different conditions. If the number unattempted is basically highly and positively correlated with the number of wrong answers, then those people with the slowest rates of work show little promise of improving their position if they are

given more time. On the other hard, if the correlation is high but negative, there is reason to believe that the slowest people are the very ones who would end up with the top scores if they were given sufficient time.

Since Gulliken's two ratios do not consider this correlation, they tend to be most meaningful in the extremes and not for the typical test. Cronbach and Warrington (1951) offer an approach which focusses on this essential correlation. They advance a formula which calls for determining the correlation between two parallel forms of a test under two conditions: with a time limit and without. If this correlation is low, then not having a time limit makes a big difference, and the test is speeded. As the correlation approaches a higher positive value, the test is increasingly less speeded, even though there may have been a marked shift in the values of Gulliken's ratios.

The basic difficulty with the Cronbach and Warrington approach is the requirement for two parallel forms and two administrations. It is more often the case that speededness needs to be evaluated in connection with a test for which there has been only a single administration. At the present time there is no single adequate index for performing the evaluation. Stafford (1971) has suggested an index, the Speededness Quotient, which is the ratio of the number of unattempted items to the total number of items which were not succeeded upon. Not succeeding on an item may be due to making a wrong answer, to omitting the item but moving on to a later one,

or to a failure to attempt. His index has the virtue of simplicity and of providing a measure of the degree of test speededness for a specific individual or a specific item. It has the basic flaw, however, of ignoring the correlation between rate-of-work and success on items. Further, by focussing on the averages of such indices as items attempted, items answered wrongly, etc., rather than the variances, as Gulliksen does, Stafford's index seems weaker. It is the differences among persons which are of interest, and the consequences of these differences for the score distribution. A focus on the average values of distributions may pose logical difficulties in interpretation. If for example, all the examinees have the same rate of work, they may on the average fail to complete a considerable number of the items. But the score variance is due to variations in their ability to handle the problems, and the test is basically a power test. Gulliksen's index, focussing as it does on variance, would be sensitive to this and would characterize the test as power. Stafford's index, focussing on averages, would not.

Educational Testing Service has long evaluated three characteristics of the completion activity of the population taking the test: (a) the percent completing the test, (b) the percent completing 75% of the test, and (c) the test item at which approximately 80% of the total group are still working. These data are combined judgmentally as criteria which make a test speeded if (1) fewer than 100% of the candidates reach 75% of the items and (2) fewer than 80% of the candidates finish 100% of the items. As Swineford (1956) observes,

however, "These are arbitrary criteria and should not, of course, be too strictly applied.... It is important to understand,....that if the criteria are not met, it does not necessarily follow that the test is speeded." Additional data is prepared, in the standard test analysis, which can help the test evaluator to a judgment. This additional data centers on the number of scores which are in the range of chance and the number of scores which seem to reflect a high level of unattempted items but not a lack of success on those attempted. In addition, the distributions of four measures: number Right, number Wrong, number Omitted and number Not Reached are presented and summarized. (An Omitted item is one which is not marked but which is followed by a mark for some later item; Not Reached item is not marked but is not followed by any later marks.)

The ETS criteria are reviewed in a discussion of test speededness in each test analysis. No single index or guideline has been developed. Evans and Reilly (1972) used these criteria for speededness but introduced a graphic technique which plots the percent of candidates who are still working at various points in the test. In their presentation, they used a base line which was the number of items, ranging from zero to the total number in the test. A more general approach simply plots the "percent of subjects still working" as a function of "percent of test worked on." An example would be:

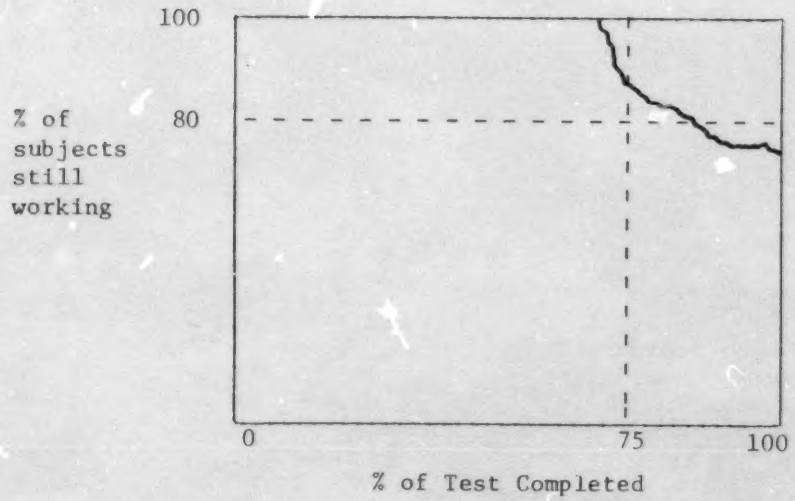


Figure 1

Such diagrams will characteristically exhibit the picture of a square with a "chunk" missing in the upper right hand corner. This is so because there is seldom any decrease in the percent of subjects still working until they begin to approach later items, and "it is seldom the case that the line of decreasing percent working" drops all the way to zero. Speed may seem to be a significant factor, but usually some people finish the test, and they constitute a sizable percentage of the subjects.

Figure 1 includes dotted lines which describe the basic ETS criteria. The example depicted is a test which would fail to meet both of these criteria; a few people fail to complete 75% of the test, and fewer than 80% of the subjects are still working on the last item. Only tests for which the graphic plot lies entirely

within the upper right area defined by the dotted lines will meet the ETS standards.

Of course, the lines can be considered for specific item effects as well as for overall test speededness. The plot of Test A in Figure 2 would indicate a basically unspeeded test but one on which the last items, for some reason, discouraged candidates from making an attempt. The failure to meet the formal criteria would need to be tempered by a consideration of the precipitous decline in the percent working. Most considerations of test speededness assume that there are no such glaring impacts by individual items.

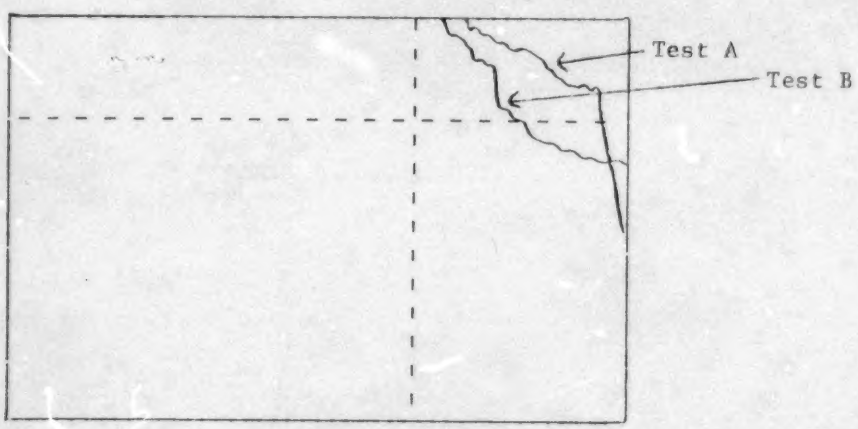


Figure 2

The plot of Test B in Figure 2 would similarly reveal that there was a single item which seemed to stop a significant number of subjects dead in their tracks. Compared to the impact of the

items which preceed and follow it, this item caused a much greater slowing down. The consideration of the graph would then permit the judgment that the test, without this particular item, would perhaps show much less evidence of speed, might in fact meet the stated criteria.

The demonstration of such aberrant items confirms the wisdom of Swineford's comment. No single general index of test characteristics will be free from errors of interpretation due to the presence of idiosyncratic items. Graphic presentations such as those in Figures 1 and 2 must be inspected before criteria used in evaluating them are applied. Nonetheless, there is a need to develop better ways of conveying the characteristics of test speed, and the route to such better ways would seem to lie in the development of better descriptors of the characteristics of the line which graphs the decrease in percent working. The ETS standard criteria simply ask, in effect, whether the line is found in its entirety within a single bounded region. This incorporates very little of the information which is available.

The graph does not lend itself readily to translation to the framework of Gulliksen or Stafford. Their ratios consider the wrongness or rightness of responses, while the graphic approach considers only items unattempted. Nonetheless, a pure power test, in Gulliksens sense will be a test with no descending line at all: 100% of the subjects are still working on 100% of the test. A single point in the field meets this criterion. A pure speed test,

on the other hand, will have a graph which fails to reach the right-hand boundary, which instead "bottoms out", with 0% of the subjects reaching the higher percentages of the test. Figure 3 presents these situations.

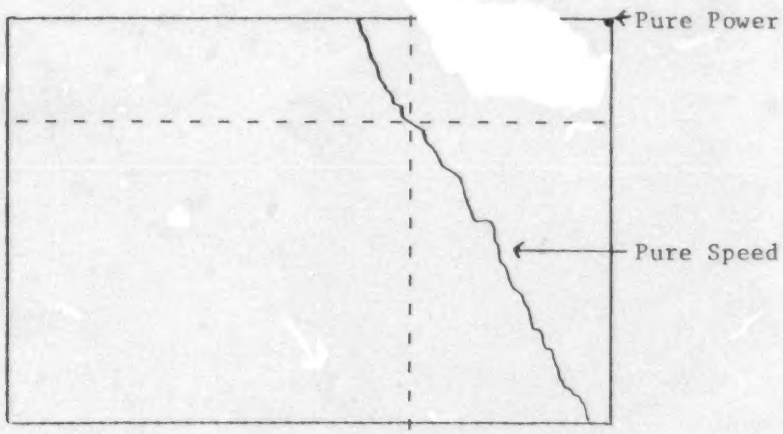


Figure 3

Further, since the graphic method provides no direct estimate of the correlation between speed and poser, the kinds of information it can yield will inevitably be partial and will rely upon assumptions for their applicability. The graph of a pure speed test in Figure 3, for example, will only be legitimately interpretable in this way if there is 100% success on the items which are attempted by the candidates. It cannot be inferred from the data on percent attempting whether this is the case or not.

It is useful, however, to explore the implications of the graphic method and to redefine the baseline in terms of rate of work.

That is, for any test for which a time limit is established for all candidates, it is possible to convert the baseline into a scale of items per minute. Thus, if the test is a 100-item test, given in 50 minutes, all those who complete the test work at a rate of 2 items per minute or greater. It is not possible to establish, for a given individual who finishes, whether this person's rate is greater than 2.0 but the lower bound for the rate is establishable. For a candidate who completes 90 items in 50 minutes, however, the actual rate is more readily estimated: 1.8 items per minute. Similarly, 80 items in 50 minutes would indicate a 1.6 items per minute rate. Thus, for any candidate who fails to finish the test, but who is still working at the end, we can establish an estimate of rate of work.

Using these estimated rates of work, we can develop projections of the required additional time that individuals might require in order to complete the test. For example, if the subject has completed 90 items in 50 minutes, working at a rate of 1.8 items per minute, then in order to complete the test, to answer the remaining 10 items, we would estimate that the subject needs $10 \div 1.8 = 5.6$ more minutes (approximately). The average of such estimates, over all the subjects who fail to complete a test, may be useful to the test constructor, as an estimate of the additional times required to halve the number of non-finishers. It is further possible to consider this "average time required to permit non-finishers to complete the test" as a percentage of the present time allotment, and to treat this

as an index of test speededness. Thus, if a test would require only 5% more time to permit half of the non-finishers to complete, it is more of a power test, whereas a test which would require 50% more time to achieve this goal is more of a speed test. It is doubtful that such a quantity would be useful as an all-purpose index for characterizing tests as speeded or unspeeded. However, it might be of considerable value in helping the test constructor to establish new time limits for any test which is inappropriately speeded.

When the baseline is redefined in this way, as a rate of work continuum, it is possible to replot the graph in order to show the percent of the group who exhibit the various rates of work. That is, the differences in percent still working from one baseline point to another shows the percent of subjects who could maintain the slower of the two work rates but could not attain the faster. Returning to our example of a 100 item test in 50 minutes, if 85% of the subjects answer 90 items, but only 81% of the subjects answer 91 items, then 4% of the subjects have work rates greater than 1.80 items/minute but less than 1.82 items/minute. Thus, it is possible to determine the empirical distribution of work rates, for those who fail to complete the test.

A focus on rate of work indices would seem to be superior to the attention which has more typically centered on the number unattempted. For a given test, of course, the two values are

consistently related. But it is possible to consider the items per minute index as independent of any test, and to extrapolate to extimates of candidate performance on other tests, differing both in total number of items and the time limit. Because of this, the items per minute index would seem to offer significant advantages as a fundamental indicator of speed on tests. In particular, the estimation of rate of work indices and rate of success indices opens the way to a possible estimate of the speed -power correlation, as Reilly (1974) has shown.

A major limitation of the analysis thus far lies in the fact that it presents rate of work data only for those who fail to complete the test, offering only a lower-bound estimate for all those who complete. However, if the shape of the overall distribution can be assumed, it may be possible to calculate its parameters from the data available on those who fail to complete.

It will frequently be reasonable to assume that the rates of work are distributed normally in the test subject population. Given this assumption, one can estimate the mean and standard deviation of the normal distribution, for the mean and standard deviation for those who fail to complete the test are known both empirically and theoretically. That is, if 35% of the total group fail to complete the test, their average rate of work and the standard deviation of these rates can be calculated empirically. Further, the mean and standard deviation of the z-scores of the hypothetical normal curve can be calculated. By combining these as linear equations, it is

possible to solve for the estimated mean and standard deviation of the rates of work of the total group.

The following formulas are used to calculate the mean and standard deviation of the z-scores of these not finishing, on the assumption that they are a subgroup of a normal population.[1]

$$\mu'_x = -\frac{1}{(p)\sqrt{2\pi}} \int_{z_p}^{\infty} x e^{-x^2/2}\, dx \qquad (1)$$

$$\sigma'_x = \left[\frac{1}{(p)\sqrt{2\pi}} \int_{z_p}^{\infty} x^2 e^{-x^2/2} dx - \mu'^2_x \right]^{1/2} \qquad (2)$$

μ'_x = the mean z-score for the group who fail to finish

p = the proportion of the sample who fail to finish

z_p = the z-score above which lies a proportion of area under the normal curve equal to p

σ'_x = the standard deviation of the z-scores of those who fail to finish.

---

[1]The author is grateful to Richard Reilly for describing this approach to calculating the parameters.

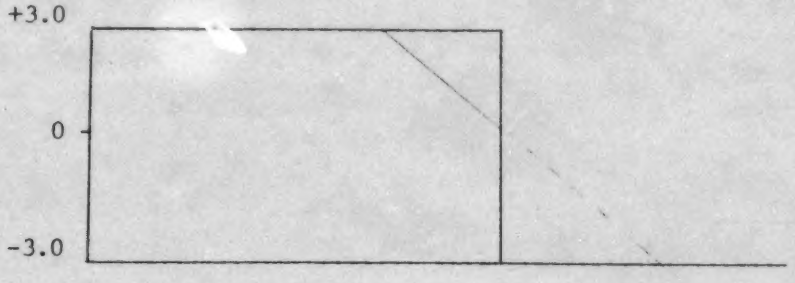
When σ'_x and μ'_x are known, the mean and standard deviation of the total group can be calculated as follows

$$S_y = S'_y/\sigma'_y \tag{3}$$

$$\bar{Y} = \bar{Y}' - S_y x \quad . \tag{4}$$

The calculation of the parameters permits a description of the rates of work for the entire population, not just those who do not finish. While specific individuals or subgroups among those who finish cannot be identified, probability statements can be developed as to the frequency with which work rates will be exhibited.

Given the estimated parameters for the total group, it is possible to replot the graphs for test speed by changing the ordinate to a scale from + 3.0 to -3.0. These z-scores values will encompass virtually the entire sample, and may be taken as effective replacements for the percent finishing. The resulting plot would be as follows:



Rate of Work

Figure 4

The line relating z-scores in the observed distribution to rates of work is presented in two segments. The solid portion is the observable segment; it describes the relationship for those who fail to complete the test. The dotted line is not observable, since all that is known about those who complete the test is a lower bound. The mean rate of work for the total sample is the point on the base-line which lies vertically beneath the intersection of a horizontal drawn from 0 on the abscissa and the sloping line.

In theory, for parallel sets of material, this line will not shift. A test will span a certain segment of possible rates of work, but the effect of changing the test characteristics, in terms of time limits or number of items, is simply to alter the proportion of the sloping line of relationship which is observable or not. The pure power test will have a sloping line which is entirely in the dotted line region. The pure speed test will have a solid line, with the normal distribution of the work rates entirely accessible to empirical confirmation.

The analysis of the characteristics of this line is an extension of current practice in the description of test speededness. While no true index of speed can avoid the consideration of speed-power correlation, a consideration of the line and its properties can be helpful in visualizing the total impact of rate of work on a test. In a sense, under the present formulation, approaches to speed such as those practiced at ETS are seen to place heavy emphasis on minimizing the solid portion of the line. The criteria of completion

(100% of the test by 80% of the group, 75% of the test by 100% of the group) establish implicit characteristics for the location of the line with respect to the rates of work sampled by the test. But the slope of the line itself does not enter into the evaluation, provided that the line is constrained to the acceptable zone.

The explicit development of the line permits a consideration of such factors as time to finish early and reexamine one's work. Examine the two tests described in Figure 5. Each is essentially unspeeded in that virtually all of the samples complete all of the tests. But Test A has a steeper slope, and this implies that there

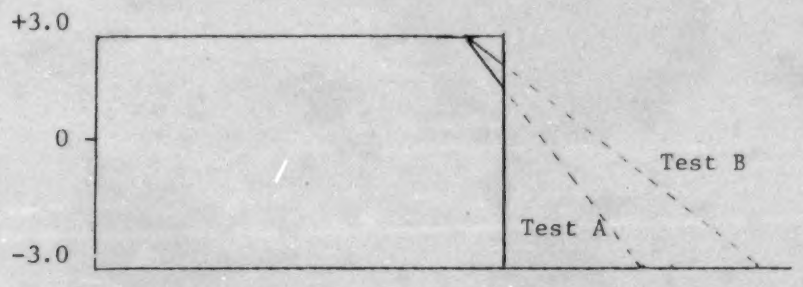


Figure 5

is a much narrower range of work rates in the group for this test. The result is that some of the hidden characteristics of test speed, such as the opportunity to reexamine work, are less involved in Test A than they are in Test B.

Such slopes will, of course, influence the correlation between rate of work and rate of success, for the diminished variance in rates of work will lower the potential correlation. They will also

figure in such aspects of tests as the extent to which subgroups of the population finish early and become restive, disturbing other subjects who may still be working, and posing problems for proctors. While no total generalization is possible, tests which have steeper slopes will be superior to other tests which have a greater spread of rates of work.

It should be explicitly pointed out that since all test analyses will have a common abscissa, ranging from +3.0 to -3.0, the slope parameter is in fact an index of the range of work rates observed in the population. The line is selected graphically for its advantages in clearly providing comparisons of tests and in clearly demarcating the zones of empirically determined and theoretically determined data. But a different approach might express many fo the same concepts in terms of the ranges of rates of work which are observed. As will be seen below, the use of range concepts is advantageous in discussions of possible indices.

The development of the concept of the line suggests several indices for test speededness.

1) A first index might be simply the total range of rates of work. In general, a test is less speeded if this range is smaller (i.e., if the slope of the line is steeper.) This will be true regardless of where the range of rates sampled by the test may lie. If the range sampled by the test clearly exceeds that exhibited by the sample, so that all

of the line is solid, nonetheless a very steeply sloped line will indicate a test for which there is little variance due to speed.

2) A second index might be the proportion of the total range of rates of work in the group which overlaps with the range of rates of work which is sampled by the test. This is again related to the slope concept but has a somewhat different set of applications. It could not be calculated for pure speed or pure power tests, which have totally overlapping or totally nonoverlapping ranges between tests and group. For the intermediate cases, however, it would provide a measure of speed. In general, the smaller the proportion of overlap, the less the test is speeded. In a sense, this index provides a possible metric or scale between the polar concepts of speed and power as suggested by Gulliksen's formulation.

The two indices described above underscore the complexity of analysis in this area, for in general, tests of different slopes, or ranges or rates are not equally sensitive to the second index. As shown in Figure 5, Test A would be considered less speeded by the first index, but more speeded by the second. It is as if the second index asked "How much overlap, in terms of rate of work, is there between test and group?" While the first index asked "How much difference can it make how much overlap there is?"

Each question is legitimate. Given two tests of equal slope, the second index will differentiate between them appropriately. Given two tests which are equal by the second index, the first index will differentiate between them appropriately. But a single number for making both of the comparisons is difficult to devise.

3) A third index might be the product of the range of rates exhibited by the group and the proportion of this range which overlaps with the range called for by the test. The greater this index, the greater the amount to speededness. How useful this single index would be in practice in combining the two dimensions of the problem would have to be decided by user experience. There is so little familiarity with the basic nature of rates of work that the numbers would not be meaningful at first.

The development of these indices proceeds from the model but there is a need to test the basic assumption of normality of the distribution. This is possible by considering the extent to which those items which are not completed by the entire group are consistent in their information. If the model is fulfilled, the z-scores which are computed from a consideration of the proportion reaching each item will be closely related to the z-scores which are predicted from the established parameters of the normal curve. That is, the parameters of the hypothetical normal curve which describes work rates are determinable by the equations given earlier. Using these parameters,

it is possible to describe a z-score for each value of the number attempted (or, conversely, not reached). If, for example, the mean work rate is 2.00 items/minute, and the standard deviation of the rates is 0.50 items/minute, then a person who works more slowly, at a rate of 1.50 items/minute, is at a z-score of -1.00. But if the distribution is normal, there would be an empirical 16% of the group who have work rates as slow as or slower than this. The actual empirical values can be compared with those calculated from the parameters and their consistence used to test the hypothesis.

Figures 6-12 present such comparisons of the two different z-score estimates for given levels of Not Reaching, the parametric estimate and the empirical estimate. In each case, the parametric values are on the ordinate, the empirical values on the abscissa. The diagonal line of equality is drawn for each figure. If all the points fell along this line, the model would be perfectly confirmed. It is clear that the tests studied gave basically an approximate conformity to the model. A test such as VSA25-14, however, in Figure 10 would seem to be sufficiently well described that predictions about the rest of the distribution of work rates could be made. Several other tests approach this level of fit, also. A test such as VSA25-17, in Figure 12, would seem to be very poorly described by a single normal distribution. It's points seem to follow two linear trends, one for the region from 0.00z to -1.90z, the other from -1.90z to 2.50z. Perhaps something happened in the administration of this test, or there was some specific characteristic of the material which

make people shift their rate of work on the later items. The evidence of Figure 12 indicates that the last few items are suddenly much more time consuming than would have been predicted from a consideration of the earlier items. In reading the charts, the items at the upper right, with the higher absolute values of z-score, are the earliest items to show individuals who do not reach them. Being early, they indicate very slow rates of work. Later items, up to the last item itself, indicate faster rates of work, nearer the average of the distribution. Thus, in Figure 12 the last items show much larger groups of people in each rate-of-work category, for the shift in empirical z-score is much greater than would be indicated by a normal distribution.

Some of the plots look as though there would be a better fit if the parameters were somewhat different. For example, Figure 11, VSA13-24 seems a reascnably linear relationship. If the parameters were somewhat different, the line of equality might pass up through the swarm of points in a much more direct manner. The estimates of the parameters derived for this paper might be improved on by requiring more precision in the statistical work. Alternately, perhaps there are superior statistical techniques for deriving these estimates.

The aberration in the plot of PB01- Error Recognition items, in Figure 8, is interesting. A small set of items appears to swerve out of the plot. This swerve is interpreted to indicate that the first items in this cluster were answered more readily than average, while the last ones took somewhat longer. This swerve is in a sense a

microcosm of the larger curve of Figure 12. It would be interesting to consider the specific properties of these items to see if reasonable hypotheses can be developed for explaining their speededness characteristics.

Figure 13 is reprinted from an earlier paper Donlon (1973). In that earlier paper the focus was on establishing time limits for tests, and graphic estimates were used to derive the parameters. For each item, the percent not reaching it was converted to a z-score, using tables of the normal curve, and these points were plotted. A line of best fit was drawn through the points by hand, and the mean and standard deviation of work rates for the total distribution estimated in this way. Thus, in Figure 13, for VSA25-13 the values of 39 and 9 were estimated, 39 because this was the value on the abscissa corresponding to a z-score of 0.00, and 9 because this was the approximate shift in abscissal values for a shift of 1.00 units on the ordinate. The values estimated for these parameters by the equations were 38.41 and 8.41, respectively, as can be seen in Table 1.

Table 1 compares the graphical and statistical estimates for six of the seven tests presented in the figures. VSA25-17 was too non-normal to make a meaningful estimate statistically. In general, the estimates agree sufficiently that the establishment of time limits by the two methods would be reasonably consistent.

Insert Table 1 about here

In summary, this paper has reviewed the use of indices of speededness in tests and has proposed three possible indices which derive from a model which assumes that rate of work is normally distributed. Each of these indices is seen as limited by the failure to adequately consider the correlation between speed and power, but they have the advantage that they are derivable from a single administration of a test. The plausibility of an assumption of a normal distribution of work rates was tested on empirical data from seven tests, and in six cases the assumption was not unrealistic although the fit was only approximate. The analysis of tests for which the assumption is not completely tenable will often be instructive for those who construct instruments. Specific item material seems to govern departures from normality.

The goal of a single, intuitively satisfying index of speededness in tests is important. Without some sort of metric or scale with which to assess degrees of speededness, the evaluation of tests in this important area remains inordinately subjective.

References

Cronbach, L. J., & Warrington, W. G. Time limit tests: Estimating their reliability and degree of speeding. Psychometrika, 1951, 16, 167-188.

Donlon, T. F. Establishing appropriate time limits for tests. Presented at the Fall, 1973 meeting of the Northeast Educational Research Association.

Evans, F. R., & Reilly, R. R. A study of speededness as a source of test bias. Journal of Educational Measurement, 1972, 9 (2), 123-131

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Reilly, R. R. Estimating total score and item completion statistics when all examinees do not finish. Presented at the Spring, 1974 meeting of the National Council on Measurement in Education.

Stafford, R. E. The speededness quotient: A new descriptive statistic for tests. Journal of Educational Measurement, 1971, 8 (4), 275-278.

Swineford, F. Technical manual for users of test analyses. Statistical Report 56-42. Princeton, N.J.: Educational Testing Service, 1956.

Table 1

Comparisons of Graphic and
Statistical Estimates of Parameters

	Graphic	Calculation
VSA25-14	32	32.38
	8	7.92
VSA25-13	39	38.41
	9	8.41
VSA13-24	73	70.24
	15	12.15
VSA25-17	No Estimate[2]	
PBO1 Usage	41	44.18
	8	9.79
PBO1 Error	45	44.24
Recognition	11	9.62
PBO1 Construction	30	29.62
Shift	10	8.26

[2]The data in this graph were too erratic to provide a plausible estimate.

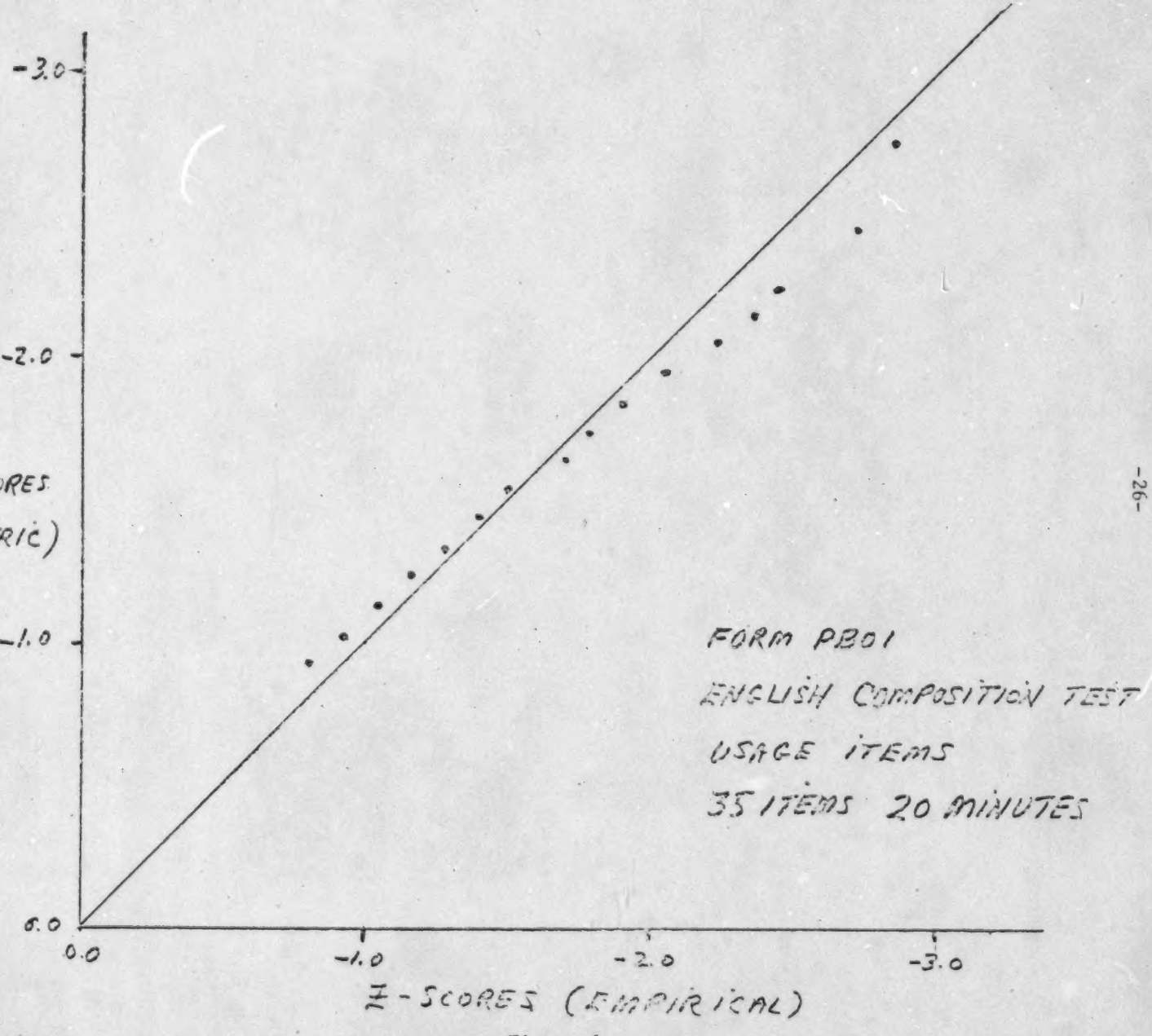

-3.0
-2.0
-1.0
0.0
Z - SCORES
(PARAMETRIC)
0.0
-1.0
-2.0
-3.0
Z - SCORES (EMPIRICAL)
FORM PB01
ENGLISH COMPOSITION TEST
USAGE ITEMS
35 ITEMS 20 MINUTES

Figure 6

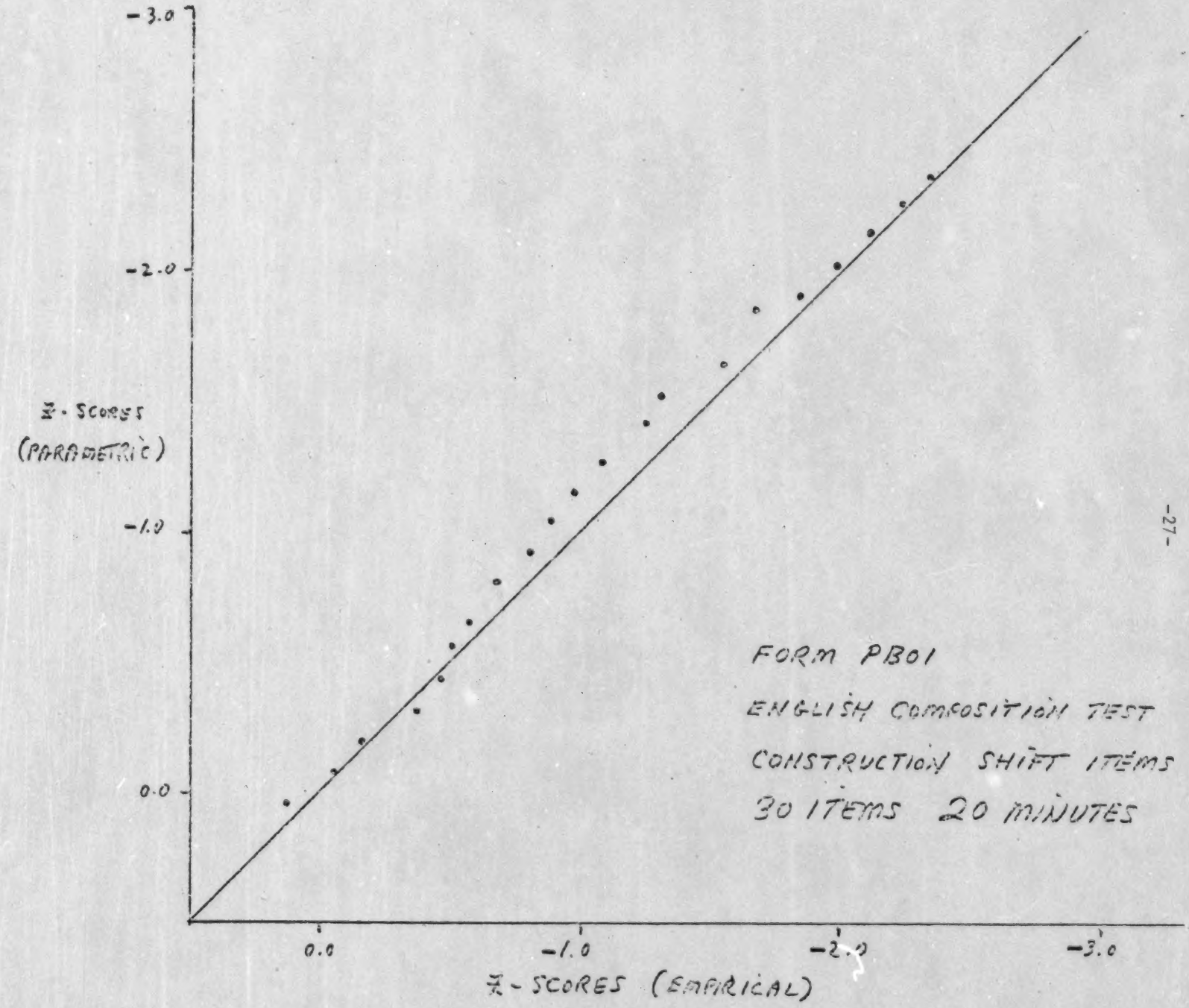

-3.0
-2.0
-1.0
0.0
Z - SCORES
(PARAMETRIC)
0.0
-1.0
-2.0
-3.0
Z - SCORES (EMPIRICAL)
FORM PB01
ENGLISH COMPOSITION TEST
CONSTRUCTION SHIFT ITEMS
30 ITEMS 20 MINUTES

Figure 7

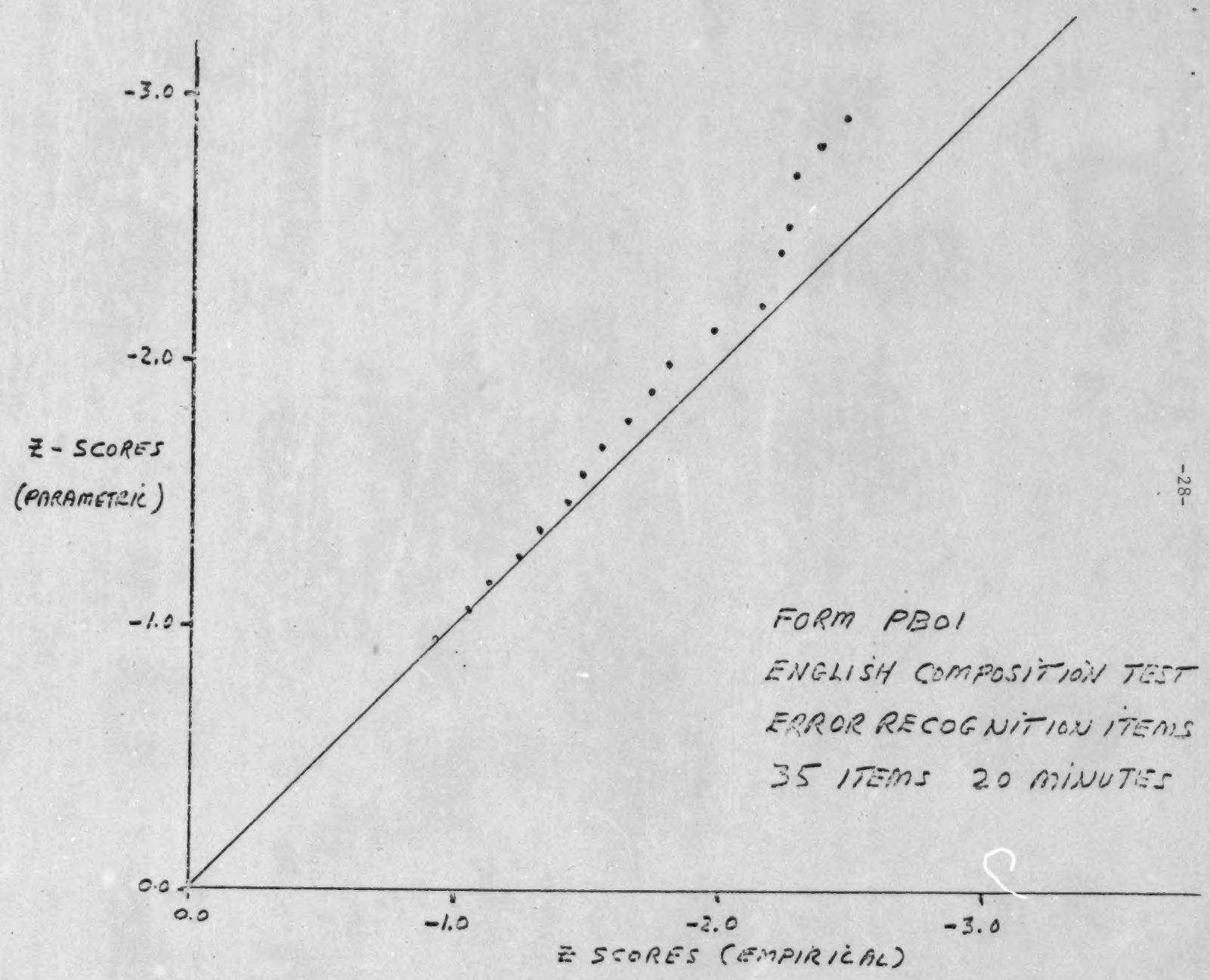

-3.0
-2.0
-1.0
0.0
Z - SCORES
(PARAMETRIC)
0.0
-1.0
-2.0
-3.0
Z SCORES (EMPIRICAL)
FORM PB01
ENGLISH COMPOSITION TEST
ERROR RECOGNITION ITEMS
35 ITEMS 20 MINUTES

Figure 8

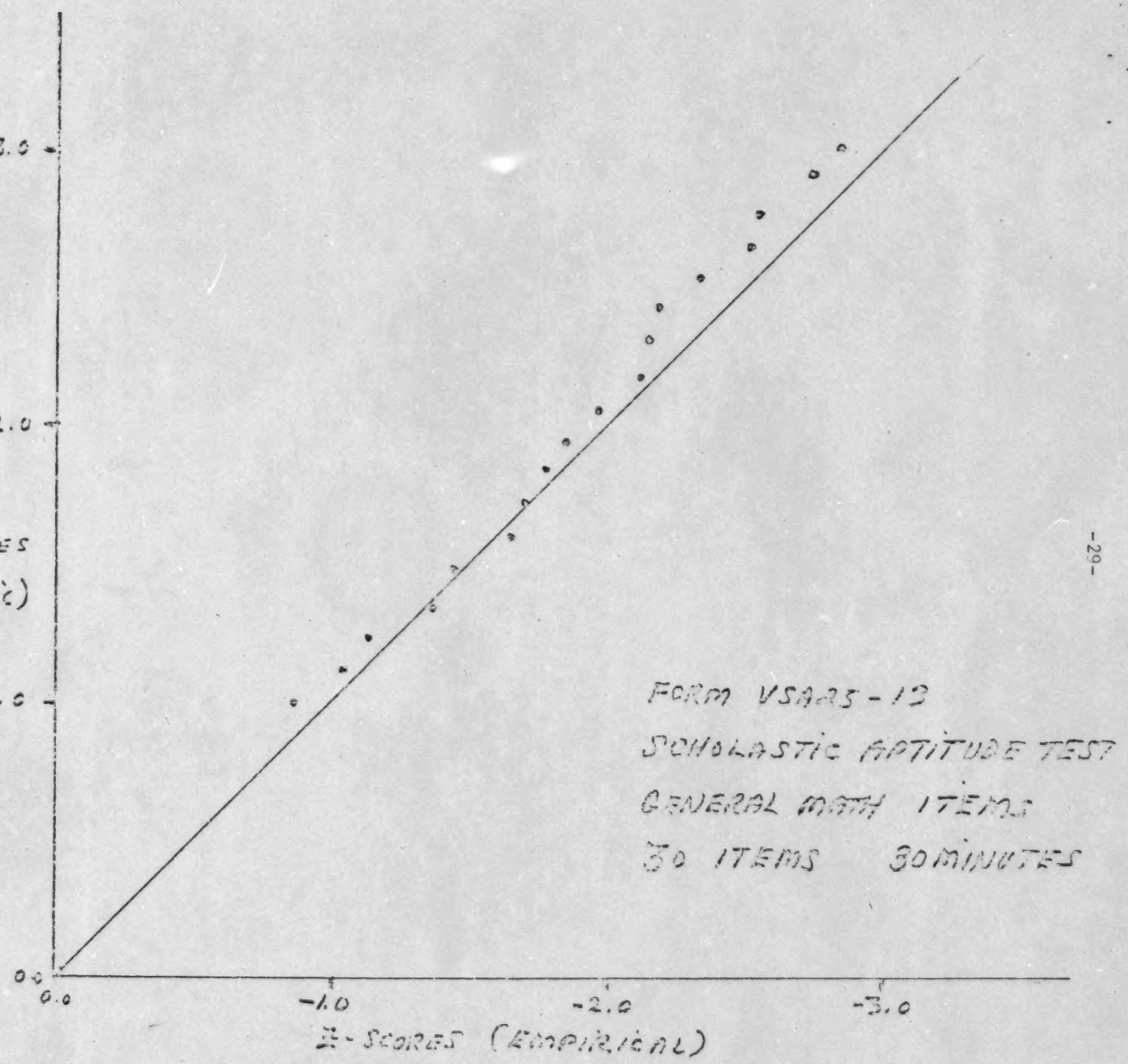

-3.0
-2.0
-1.0
0.0
0.0
-1.0
-2.0
-3.0
Z - SCORES (PARAMETRIC)
Z - SCORES (EMPIRICAL)
FORM VSAAS-12
SCHOLASTIC APTITUDE TEST
GENERAL MATH ITEMS
30 ITEMS 30 MINUTES

Figure 9

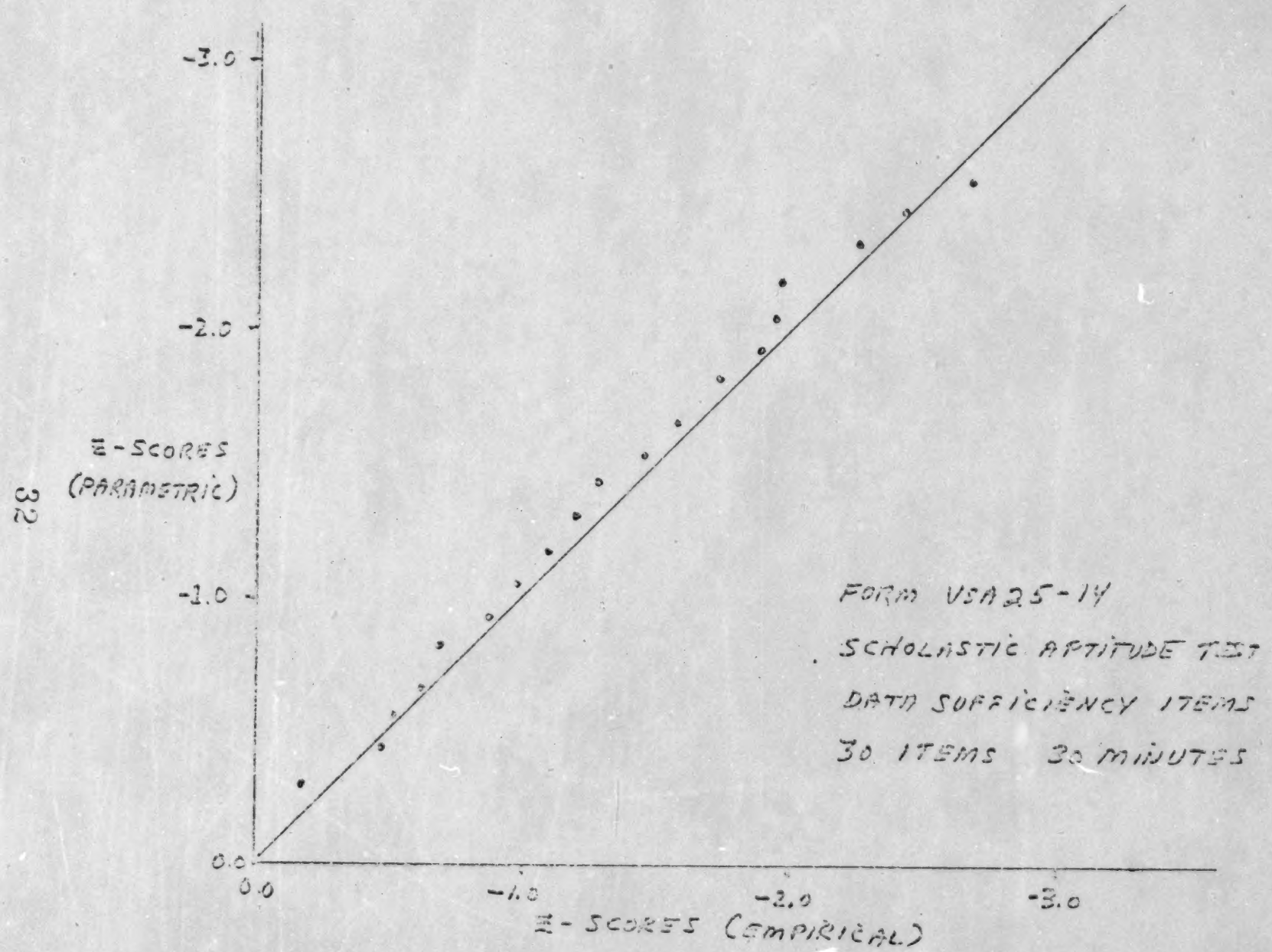

-3.0
-2.0
-1.0
0.0
Σ-SCORES (PARAMETRIC)
0.0
-1.0
-2.0
-3.0
Σ-SCORES (EMPIRICAL)
FORM VSA 25-1Y
SCHOLASTIC APTITUDE TEST
DATA SUFFICIENCY ITEMS
30 ITEMS 30 MINUTES

Figure 10

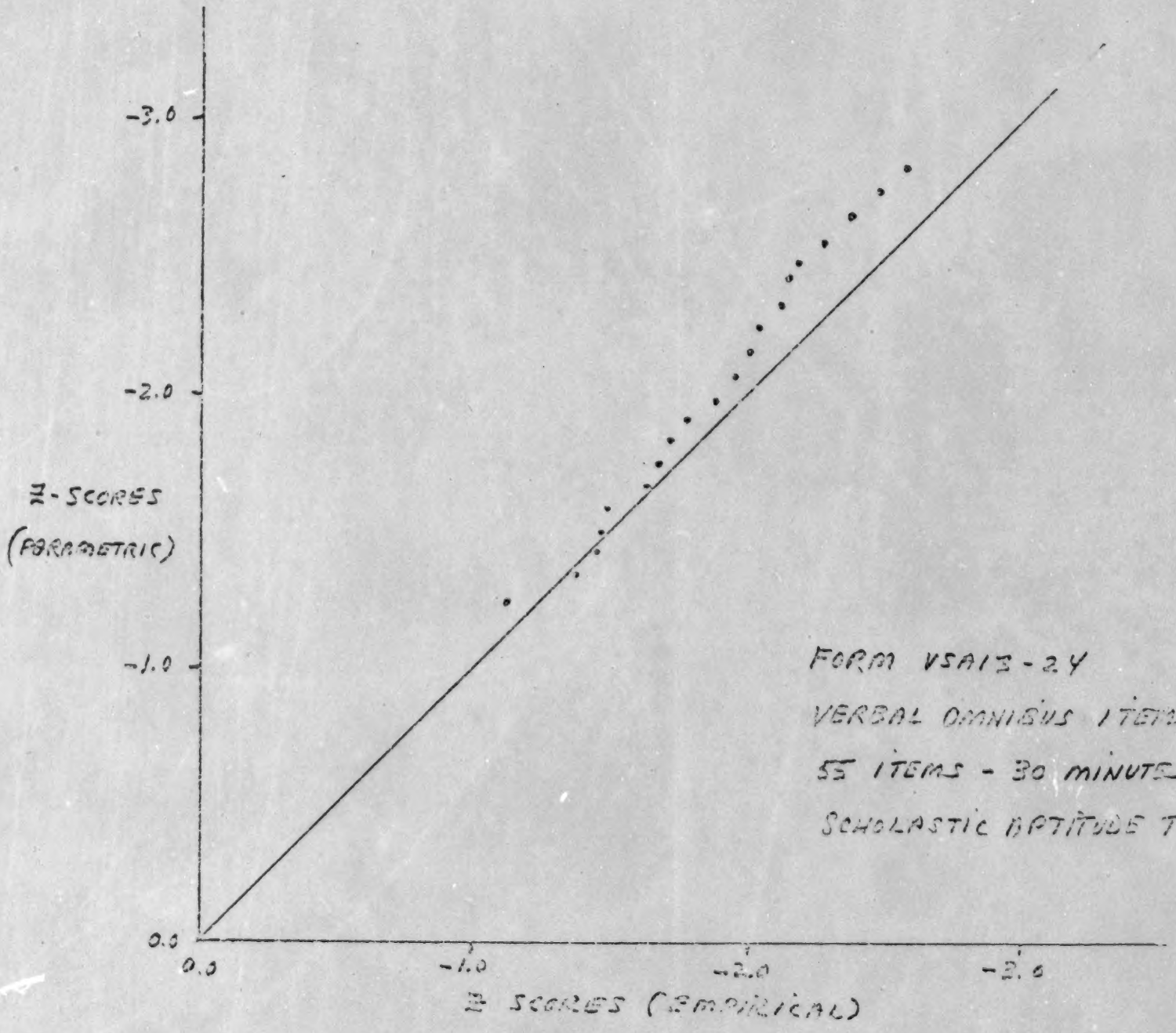

-3.0
-2.0
-1.0
0.0
Z-SCORES
(PARAMETRIC)
FORM VSAIS-24
VERBAL OMNIBUS ITEMS
55 ITEMS - 30 MINUTES
SCHOLASTIC APTITUDE TEST
0.0
-1.0
-2.0
-3.0
Z SCORES (EMPIRICAL)

Figure 11

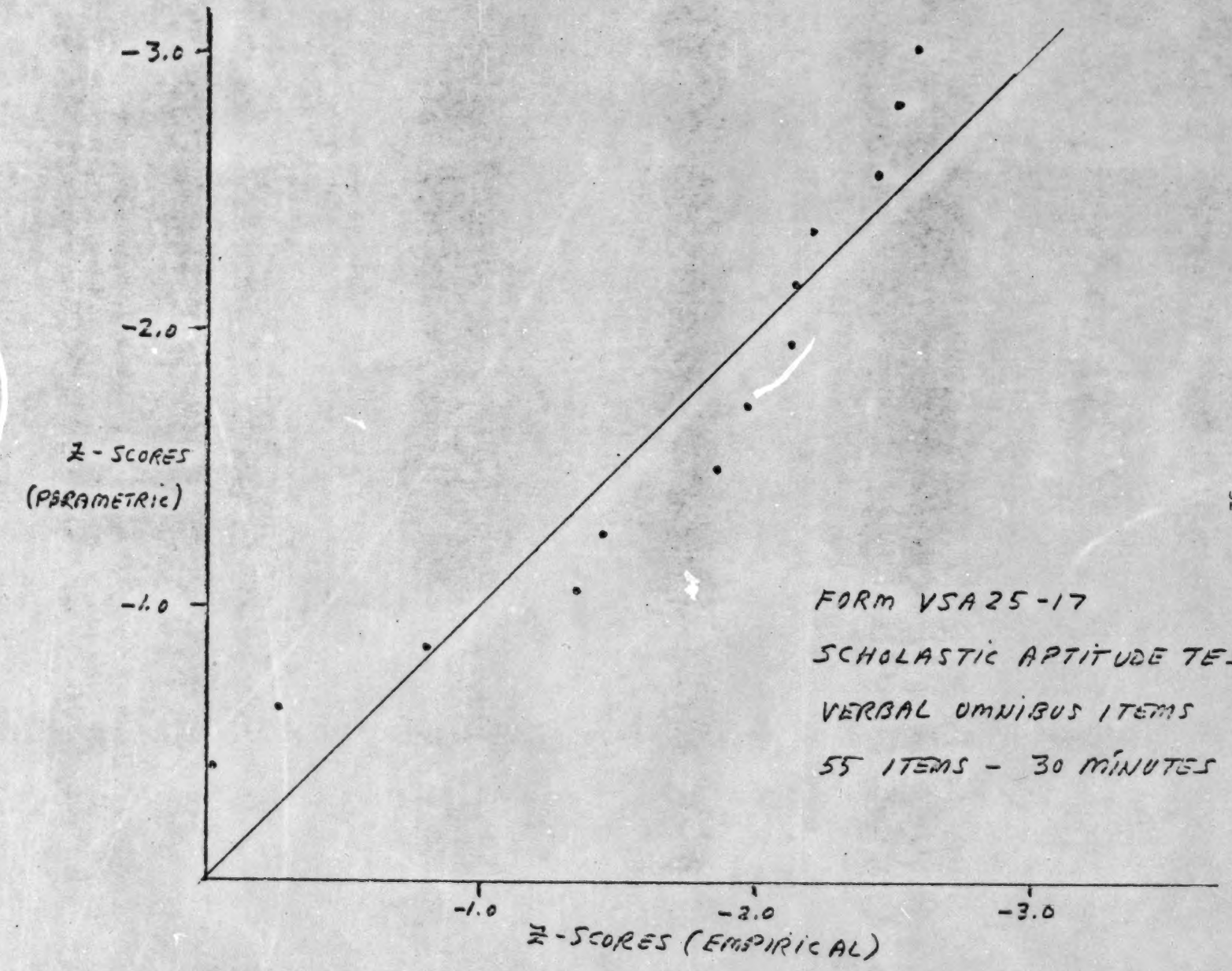
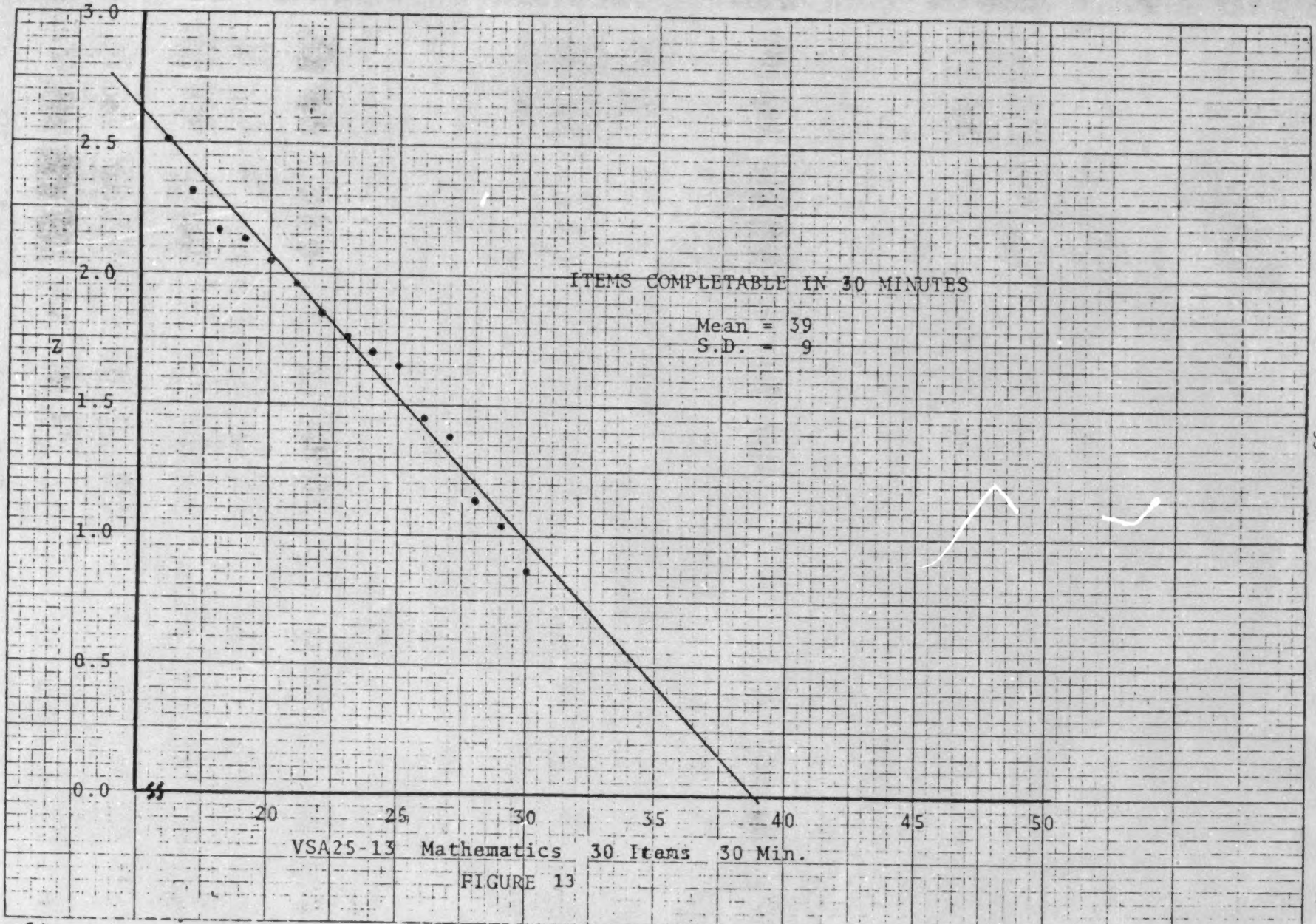


Figure 12

BEST COPY AVAILABLE



VSA25-13 Mathematics 30 Items 30 Min.

FIGURE 13

BEST COPY AVAILABLE